DOCUMENT RESUME

ED 061 296 TM 001 295

AUTHOR Proper, Elizabeth C.
TITLE Review of Problems of Testing for Homogeneity Prior to Running an ANOVA.
PUB DATE Jun 71
NOTE 13p.; Paper presented at the Annual Conference of the New England Educational Research Organization, Chestnut Hill, Massachusetts, June 1971

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Analysis of Variance; Hypothesis Testing; *Literature Reviews; Models; *Research Methodology; Statistical Analysis; *Testing Problems; Tests of Significance

ABSTRACT

Texts often suggest running preliminary tests for homogeneity of variance prior to running an ANOVA. While it has been known for some time that most of the suggested tests are probably not appropriate, they are still being used. This paper is a review of the literature in terms of the implications involved in running preliminary tests in general and various ones in particular: Cochran, Hartley, Box and Andersen, Bartlett, Levene. It re-emphasizes the need to attain equal cell sizes and suggests the appropriateness of the Welch test when that is not possible. The paper looks at the difference in assumptions which must be met in the fixed and random effects models, in a one-way design. (Author)

ED 061296

REVIEW OF PROBLEMS OF TESTING FOR HOMOGENEITY PRIOR TO RUNNING AN ANOVA

Elizabeth C. Proper
University of Massachusetts

TM 001 295

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

(Paper presented at NEERO Annual Conference, Chestnut Hill, Mass., June 1971.)

An assumption underlying the analysis of variance is homogeneity of variance. This paper is concerned with the random and fixed effects models in the one way design. It discusses briefly the scope and limitations of various tests which have been suggested as preliminary tests as well as other tests which are available but are not often suggested. An examination is then made of the appropriateness of the use of such a preliminary test.

While most statistics books deal with the problem of homogeneity, they treat it at varying levels of significance and thus provide responses to it and suggestions for dealing with it at differing levels. This paper will point out problems involved and will identify various inadequacies in solutions which have been proposed for handling the assumption. The problem is important because homogeneity is a potential issue every time that one uses analysis of variance. The tests to be discussed are the Hartley, Cochran, Bartlett, Wald sequential analysis, Bartlett and Kendall, Box and Andersen, and Levene.

The test proposed by Hartley, the F_{max} test (Winer, 1962),

$$F_{max} = \frac{\text{largest of k treatment variances}}{\text{smallest of k treatment variances}}$$

perhaps the easiest test to compute, has as its parameters: n-1 degress of freedom and the k treatments; a special table exists for interpretation. The n in the degrees of freedom is the sample size within each treatment group; therefore, the test requires equal n's. In a case of slight inequality, the largest of the n's may be used; this will result in the null being rejected more often that it should be; in other words, such usage will result in a slight positive bias.

Winer (1962) points out that this test makes use of what is equivalent to the range of the sample variances. Making use of only two pieces of data, it is not a sufficient statistic. The problem of

sensitivity to non-normality of the F_{max} test will be handled in conjunction with the other tests.

The Cochran test (Winer, 1962),

$$C = \frac{\text{largest of k treatment variances}}{\Sigma \text{ treatment variances}}$$

as the F_{max}, uses as its parameters, the k treatment groups and n-1, the degress of freedom for each of the treatment variances, thus depending upon equal groups, and resulting in a slight positive bias when the largest n of slightly unequal groups is used. There is a special table for interpretation. Because this test uses more information, is more sufficient, it is more sensitive than the F_{max} test; the practical import of this fact, however, is negligible in terms of the concerns in this paper as will be noted in the general discussion of sensitivity to non-normality and appropriateness of preliminary testing. However, conceptually, it appears that this statistic is still insufficient in that the variability of variances other than the largest, which is used as the numerator in the equation, is taken into consideration in only a secondary way, through examination of their totality; it is conceivable that a more sufficient statistic would take into account the proportion and/or distribution of the k-1 variances. For example, two different five cell designs might have the same largest within cell variance and the same total variance, but the other k-1 variances in the two designs might be quite different. 10,10,1,1,1 and 10,4,2,4,3 would each have a largest within cell variance of 10 and a sum of within cell variance of 23, but they do not share similar proportionality or distribution of variance.

The Bartlett test (Winer, 1962) is

$$\chi^2 = \frac{2.303}{c} (f \log MS_{error} - \Sigma f_j \log s^2_j)$$

where

$f_j = n_j - 1 = df$ for s^2_j

$f = \Sigma f_j = df$ for MS_e

$$c = 1 + \frac{1}{3(k-1)} \left(\Sigma \frac{1}{f_j} - \frac{1}{f}\right)$$

$$MS_{error} = \frac{\Sigma SS_j}{\Sigma f_j}$$

k = no. of treatment groups.

The χ^2distribution with k-1 legrees of freedom to determine significance is considered by Winer (1962) to be "sufficiently sensitive" for use in those situations which require a preliminary test. However, he discourages such usage except in "a relatively few cases" [p. 95]. He considers the F_{max} and the Cochran to be adequate for most needs. The sampling distribution of the ratio: arithmetic mean/geometric mean which the Bartlett test uses has a smaller standard error than the sampling distribution of the range of the sample variances which the F_{max} test uses; thus the greater power for the Bartlett test (Winer, 1962). The fact that the test allows for unequal n's makes it more useful than the F_{max} and the Cochran; the laborious calculations involved discourage its use.

The sequential analysis method of statistical inference provides an alternative method of examining the data. Wald (1947) discusses application of this method for testing the H_0 that $\sigma^2 = \sigma^2_0$ when H_1 is that $\sigma^2 = \sigma^2_1$. The test assumes known population means, but there is

a modification of the procedure for unknown population means. An initial objection to common usage of this test is the necessity of working in a new and different frame of reference than that encountered by most elementary practitioners of applied educational statistics, with the result that probably it would be ignored. A second objection is that it appears to be fairly complicated to compute. A third is that Wald discusses his system of sequential analysis in terms of a normal distribution which, as will be pointed out, makes it inappropriate as a preliminary test.

Bartlett and Kendall (1946) developed a test for homogeneity of variance which utilizes the logarithms of variance estimates. According to Box (1953) this test also depends on a normal distribution.

Box and Andersen (Box, 1953; Box and Andersen, 1955) report that in tests such as the Bartlett, Cochran, F_{max}, Wald sequential, and the Bartlett-Kendall, one compares the variation of samples with a theoretical variation rather than with another internal measure, such as is done when between and within group means are compared. This results in these tests being heavily dependent upon a normal distribution. Examples in Table 1, taken from Box (1953, p. 320), illustrate the

- - - - - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - - - - -

effect of non-normality, specifically in terms of the Bartlett test. Within a leptokurtic population (one with a peaked distribution), differences will tend to be manifest where none exist; as indicated in Table 1, a kurtosis of 1 will result in the 5% level shifting to the 11% level for a two group design and to the 17.6% level for a five

group design. Within a platykurtic population (one with a flattened distribution), differences which are real will tend not to be made manifest; a kurtosis of -1, as shown in Table 1, will depress the 5% level to a .56% level for two groups and to a .08% level for five groups. Thus, significant results obtained with these tests may just as easily indicate a non-normal distribution as lack of homogeneity. Box suggests a need to studentize the fourth moment as the second moment has been studentized for the test on means.

Box and Andersen (1955) developed a test which is a modification of the F and Bartlett tests, based on permutation theory, which provides an approximate size alpha even in cases of non-normality. Their test is based on the fourth moment; it determines a correction factor for the degrees of freedom. Their data indicate that their method is adequate for normal, rectangular and double exponential distributions. However, equal cell sizes were used.

A test developed by Levene (1960) which he proposed as an alternative to the Box and Andersen test, in part because it may have greater applicability, uses the standard analysis of variance techniques on z_{ij}, the absolute differences between x_{ij} and $x_{i.}$. In his analysis, he explored the use of z, z^2, the log of z, and $\sqrt{z}$. Z and its square behave best in his analysis, his preference being z because of ease of computation. Examining the test under normal, uniform, double exponential and a bizarre C distribution (a misrun of the double exponential), he found that his test has power comparable to the Box and Andersen, although the Box and Andersen alpha levels are slightly better. As with the Box and Andersen, his tests involved equal sample sizes.

Of the tests examined thus far, it appears that only the Box and Andersen, and the Levene are not sensitive to non-normality and that the studies of them have been for equal n's. Scheffé (1959) suggests a test to determine if there is inequality of variance, which as he points out is appropriate not for preliminary testing, with which this paper is concerned, but with those cases in which the primary concern is differences between variances.

The pupose of this paper is to examine tests which might be used to test homogeneity of variance as it is an assumption underlying the analysis of variance. Various tests have been examined which have been suggested for use as preliminary tests; now an examination will be made of the circumstances in which they might be needed.

The concern for homogeneity involves Model I, but not Model II, that is the fixed effects model, but not the random effects model. This is so because in the random effects model one assumes that there is only one distribution of errors with a given variance. The errors must, however, be independent of each other and the treatments (Hays, 1963). The random effects model is sensitive to departures from normality (Kendall, 1966).

Within Model I (fixed effects) it has been noted by Box (1954) that moderate inequality of variance does not have serious effects providing that the cell sizes are equal. For example, with three groups having n's of five if the ratio of the groups variances is 1:2:3, the probability of exceeding the 5% point is 5.58; if the ratio of the group variances is 1:1:3, the probability of exceeding the 5% point is 5.87. One, therefore, does not need to test for homogeneity in the fixed effects

model if one has equal cell sizes. The primary area of concern that remains is the case of unequal cell sizes in the fixed effects model. The tests which have been reviewed here either demand equal cell sizes or are subject to being sensitive to departures from normality.

Prior to suggesting a solution to the problem of what to do if one suspects heterogeneity of variance with unequal cell sizes when one wishes to do an analysis of variance, it might be well to stop and examine whether or not one should ever run a preliminary test, even if one exists that meets the requirements. The results of a preliminary test will depend on the power of the preliminary test. Box and Andersen (1955) suggest that the concern should be the robustness of the main test. In a sense, one is removing oneself by another step from the problem when one runs a preliminary test. In the practical world of today, where most tests are sensitive to departures from normality, this means that one could reject the null hypothesis of homogeneity of variance and therefore not run an analysis of variance or play with the data prior to running it, when actually homogeneity existed, but there was a departure from normality.

Box and Andersen (Box, 1953; Box and Andersen, 1955) suggest that the answer to the problem of possible lack of homogeneity of variance with unequal cell sizes in the one way design is the Welch criterion (Welch, 1951) which uses a weighted variance in place of the pooled variance: $\sum_t w_t(\bar{x}_t - \hat{x})^2$; $w_t = n_t/s^2{}_t$. According to Box and Andersen (1955), this modified criterion "would be expected to be insensitive to differences in groups variances (and by analogy with the

standard test) to departures from normality also" [p. 3].

The Welch criterion is

$$V^2 = \frac{\sum_t w_t (\bar{y}_t - \hat{y})^2 / (k - 1)}{\left[1 + \frac{2(k - 2)}{(k^2 - 1)} \sum_t \frac{1}{f_t} \left(1 - \frac{w_t}{\Sigma w_t}\right)^2\right]}$$

where

t = treatment

n_t = number of individuals within treatment t

s_t^2 = individual within treatment variances, estimated on df f_t (one less than number of replicates in each case)

k = number of treatments

$w_t = n_t / s_t^2$

$\hat{y} = (\Sigma w_t \bar{y}_t) / (\Sigma w_t)$

$\bar{y}_t$ = treatment mean

$\hat{f}_1 = (k - 1)$

$$\hat{f}_2 = \left[\frac{3}{(k^2 - 1)} \sum_t \frac{1}{f_t} \left(1 - \frac{w_t}{\Sigma w_t}\right)^2\right]^{-1}$$

refer V^2 to variance ratio table with df $\hat{f}_1$ and $\hat{f}_2$

Scheffé points out that computations are difficult in a weighted analysis beyond the one-way level because of the loss of orthogonality involved. In a telephone conversation, Gene Glass concurred with the author that at this point there is no way of testing the assumption of homogeneity when one has more than a one way design.

A question which remains is whether or not a person running experiments in education should be concerned with a possible lack of homogenity of variance aside from possible effects it may have on the ANOVA. It would seem that if members had been randomly assigned to

treatment groups that a lack of homogeneity of variance in a post test situation might be of importance in and of itself. Our elementary text books teach about means and about testing for the difference between means; perhaps it is a mistake to make the logical inference that because text books teach this, that this is all one may find in experimental results. A lack of homogeneity of variance may not occur very often, and considering its implications in terms of simple analyses, we may be grateful that it does not; however, possibly there should be greater emphasis on the fact that when it does occur, that it might be a treatment effect on within cell variances.

Table II is a review of those areas in which the assumption of homogeneity is important and notes ways of handling the problem. The

- - - - - - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - - - - - -

assumption of homogeneity of within cell variance is a concern when one has unequal n's in a fixed factor one way design. In this case, one should run the Welch criterion if heterogeneity is suspected; Box (1953) suggests using the Welch criterion whenever heterogeneity is suspected, including the case of equal n's.

TABLE 1

True Percentage Chance of Exceeding 5% Level in Large Samples from Non-normal Populations

γ_2[a]	No. of groups		
	2	3	5
1	11.0	13.6	17.6
0	5.0	5.0	5.0
-1	0.56	0.25	0.08

[a] γ_2, kurtosis, is a measure of normality

TABLE 2

Suggestions for Handling Assumption of Homogeneity of Within Cell Variance in a One Way Design

Model	Cell type	Suggestions
fixed	equal n's	violation not serious
	unequal n's	use Welch criterion
random	equal n's	homogeneity of within cell variance is not a concern
	unequal n's	homogeneity of within cell variance is not a concern

REFERENCES

Bartlett, M. S. & Kendall, D. G. The statistical analysis of variance-heterogeneity and the logarithmic transformation. Journal of the Royal Statistical Society Supplement, 1946, 8 (1), 128-138.

Box, G. E. P. Non-normality and tests on variances. Biometrika, 1953, 40, 318-335.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. effect of inequality of variance in the one-way classification. Annals of Mathematical Statistics, 1954, 25, 290-302.

Box, G. E. P. & Andersen, S. L. Permutation theory in the der ation of robust criteria and the study of departures from assumption. Journal of the Royal Statistical Society, Series B, 1955, 17 (1), 1-34.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart and Winston, 1963.

Kendall, M. G. & Stuart, A. The advanced theory of statistics. Vol. 3. Design and analysis, and time series. London: Charles Griffin, 1966.

Levene, H. Robust tests for equality of variances. In I. Olkin, et al. (Ed.), Contributions to probability and statistics. Stanford: Stanford University Press, 1960. Pp. 278-292.

Scheffé, H. The analysis of variance. New York: John Wiley & Sons, 1959.

Wald, A. Sequential analysis. New York: John Wiley & Sons, 1947.

Welch, B. L. On the comparison of several mean values: an alternative approach. Biometrika, 1951, 38, 330-336.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.